Correlation and Linear Models

Prof. Wells

STA 209, 2/10/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Discuss the relationship between correlation and causation
- Compare and contrast observational studies and random experiments
- Introduce linear models

Section 1

Experiments and Observational Studies

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- Modern randomized experiments are built on 5 principles:

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- Modern randomized experiments are built on 5 principles:
  1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- Modern randomized experiments are built on 5 principles:
  1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.
  2. **Randomized**: Subjects are randomly assigned into treatment / control groups to minimize influence of variables that cannot be controlled.

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- Modern randomized experiments are built on 5 principles:
  1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.
  2. **Randomized**: Subjects are randomly assigned into treatment / control groups to minimize influence of variables that cannot be controlled.
  3. **Replicable**: Methodology should be thoroughly documented so that later researchers can replicate study to verify findings.

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- Modern randomized experiments are built on 5 principles:
  1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.
  2. **Randomized**: Subjects are randomly assigned into treatment / control groups to minimize influence of variables that cannot be controlled.
  3. **Replicable**: Methodology should be thoroughly documented so that later researchers can replicate study to verify findings.
  4. **Blocking**: If variables are suspected to affect response variable, subjects are first grouped into blocks based on these variables.

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- Modern randomized experiments are built on 5 principles:
  1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.
  2. **Randomized**: Subjects are randomly assigned into treatment / control groups to minimize influence of variables that cannot be controlled.
  3. **Replicable**: Methodology should be thoroughly documented so that later researchers can replicate study to verify findings.
  4. **Blocking**: If variables are suspected to affect response variable, subjects are first grouped into blocks based on these variables.
  5. **Blind**. When possible, neither experimenters nor subjects should know whether subjects are in treatment or control group.

Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.

Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
    - Explanatory variable: nitrate content of diet.
    - Response variable: exhaustion measured by O2 saturation.

Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
  - Explanatory variable: nitrate content of diet.
  - Response variable: exhaustion measured by O2 saturation.

  - Treatment: nitrate dietary supplement (powdered beet)
  - Control: No supplement

Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
  - Explanatory variable: nitrate content of diet.
  - Response variable: exhaustion measured by O2 saturation.

  - Treatment: nitrate dietary supplement (powdered beet)
  - Control: No supplement
- It is suspected that nitrate supplements may effect professional and amateur athletes differently,
  - We are concerned that imbalance in number of pro / amateur athletes between treatment and control groups could influence results.

Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
  - Explanatory variable: nitrate content of diet.
  - Response variable: exhaustion measured by O2 saturation.
  - Treatment: nitrate dietary supplement (powdered beet)
  - Control: No supplement
- It is suspected that nitrate supplements may effect professional and amateur athletes differently,
  - We are concerned that imbalance in number of pro / amateur athletes between treatment and control groups could influence results.
- To minimize this risk, we block subjects by pro / amateur status:

Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
  - Explanatory variable: nitrate content of diet.
  - Response variable: exhaustion measured by O2 saturation.

  - Treatment: nitrate dietary supplement (powdered beet)
  - Control: No supplement
- It is suspected that nitrate supplements may effect professional and amateur athletes differently,
  - We are concerned that imbalance in number of pro / amateur athletes between treatment and control groups could influence results.
- To minimize this risk, we block subjects by pro / amateur status:
  1. Divide SRS into pro and amateur blocks.
  2. Randomly assign pro athletes to treatment and control groups.
  3. Similarly, randomly assign amateur athletes to treatment and control groups.
  4. This ensure pro/amateur status is equally represented in treatment and control groups.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
    - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
  - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
  - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
  - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
  - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.
- Experiments may be temporally impossible
  - To study whether gender influenced survival rate on the Titanic, we only have historical records to look at.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
  - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
  - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.
- Experiments may be temporally impossible
  - To study whether gender influenced survival rate on the Titanic, we only have historical records to look at.
- Experiments may not be manufacturable
  - To study whether high unemployment rate leads to presidential losses for the incumbent party, we cannot create new presidential races.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
    - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
    - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.
- Experiments may be temporally impossible
    - To study whether gender influenced survival rate on the Titanic, we only have historical records to look at.
- Experiments may not be manufacturable
    - To study whether high unemployment rate leads to presidential losses for the incumbent party, we cannot create new presidential races.
- Experiments of appropriate size may be prohibitively expensive
    - Experiments of small or moderate size often include uncontrolled confounding variables

Random Sampling vs. Random Assignment

- Statistical investigations can incorporate two sources of randomization:

Random Sampling vs. Random Assignment

- Statistical investigations can incorporate two sources of randomization:

		Assignment of Explanatory Variable		
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)	
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions might not be generalizable because of volunteer bias.
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.	

Section 2

Assessing Relationships Between Variables

Explanatory and Response Variables

- Consider the following question:

Explanatory and Response Variables

- Consider the following question:
    - Is total spending on health care higher or lower in countries with longer life expectancy?

Explanatory and Response Variables

- Consider the following question:
  - Is total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.

Explanatory and Response Variables

- Consider the following question:
    - Is total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.
    - Ultimately, these terms are labels to help keep track of which variables we suspect may affect the others. There may turn out to be no actual causal link between the two (or the link may be the reverse of what we suspect)

Explanatory and Response Variables

- Consider the following question:
  - Is total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.
  - Ultimately, these terms are labels to help keep track of which variables we suspect may affect the others. There may turn out to be no actual causal link between the two (or the link may be the reverse of what we suspect)
- Two types of data collection methods:
  1. **Observational studies**, where researchers do not interfere with how data arises.
  2. **Random experiment**, where individuals are assigned to group and a random treatment is assigned.

Explanatory and Response Variables

- Consider the following question:
  - Is total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.
  - Ultimately, these terms are labels to help keep track of which variables we suspect may affect the others. There may turn out to be no actual causal link between the two (or the link may be the reverse of what we suspect)
- Two types of data collection methods:
  1. **Observational studies**, where researchers do not interfere with how data arises.
  2. **Random experiment**, where individuals are assigned to group and a random treatment is assigned.
- Usually, only random experiments may allow researchers to conclude a causal link between explanatory and response variables.

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.
- Correlation is bi-directional: If X is correlated with Y, then Y is necessarily correlated with X

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.
- Correlation is bi-directional: If X is correlated with Y, then Y is necessarily correlated with X
- Causality can be mono-directional: It is possible for changes in X to cause changes in Y, but for changes in Y *not* to cause changes in X.

Correlation and Causation

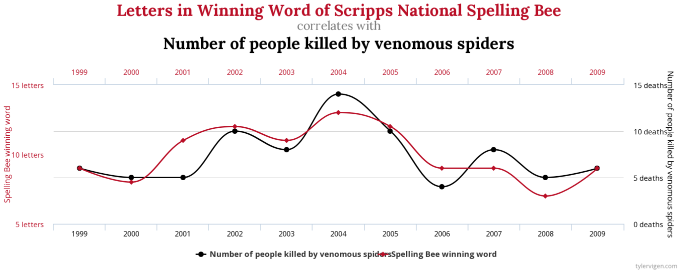
- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.
- Correlation is bi-directional: If X is correlated with Y, then Y is necessarily correlated with X
- Causality can be mono-directional: It is possible for changes in X to cause changes in Y, but for changes in Y *not* to cause changes in X.
- If variables X and Y are correlated, there are 4 possible explanations:
    1. Changes in X cause changes in Y
    2. Changes in Y cause changes in X
    3. Changes in a third variable Z cause changes in *both* X and Y
    4. The observed correlation in X and Y is due to chance.

Correlation Due to Chance

- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.

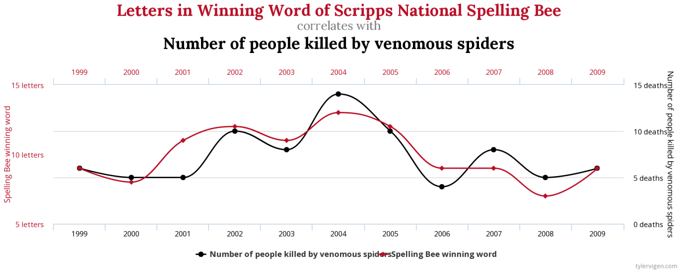
Correlation Due to Chance

- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.



Correlation Due to Chance

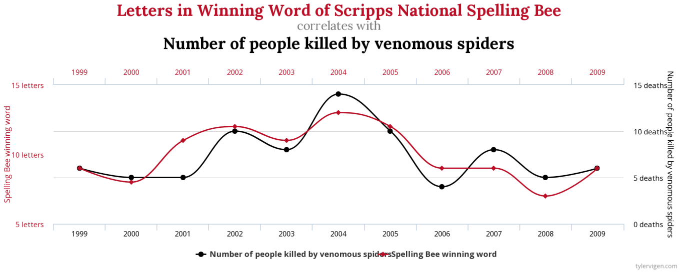
- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.



- How do we rule out spurious correlations?

Correlation Due to Chance

- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.



- How do we rule out spurious correlations?
  - Gather more data. If the correlation occurred by chance just due to sampling, the relationship is unlikely to be repeated in an independent sample.

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals
- How do we rule out confounding variables?

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals
- How do we rule out confounding variables?
  - Create models that include possible confounding variables

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals
- How do we rule out confounding variables?
  - Create models that include possible confounding variables
  - Design experiments that control for possible confounding variables

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?
  - Perhaps not. A randomized experiment involving 17,000 Belarusian infants between 1996 and 1997 found that smaller size was strongly associated with subsequent weaning and discontinuation of exclusive breastfeeding in each follow-up interval (even after adjusting for confounding variables.)

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?
  - Perhaps not. A randomized experiment involving 17,000 Belarusian infants between 1996 and 1997 found that smaller size was strongly associated with subsequent weaning and discontinuation of exclusive breastfeeding in each follow-up interval (even after adjusting for confounding variables.)
- How do you rule out reverse causation?

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?
  - Perhaps not. A randomized experiment involving 17,000 Belarusian infants between 1996 and 1997 found that smaller size was strongly associated with subsequent weaning and discontinuation of exclusive breastfeeding in each follow-up interval (even after adjusting for confounding variables.)
- How do you rule out reverse causation?
  - Investigate the temporal order of events.
  - Design an experiment where theorized cause is administered as treatment.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

  In a 1958 article in Nature, (in)famous statistician R. A. Fisher presented a case that smoking ***does not*** *cause lung cancer, arguing that:*

  "If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others."

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

  *In a 1958 article in Nature, (in)famous statistician R. A. Fisher presented a case that smoking **does not** cause lung cancer, arguing that:*

  "If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others."

  That is, according to Fisher, what if people disposed to cancer turn to cigarettes to relieve discomfort?

- Fisher did not disagree with the statistical analysis that smoking and cancer were highly correlated.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

  In a 1958 article in Nature, (in)famous statistician R. A. Fisher presented a case that smoking ***does not*** *cause lung cancer, arguing that:*

  "If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others."

  That is, according to Fisher, what if people disposed to cancer turn to cigarettes to relieve discomfort?

- Fisher did not disagree with the statistical analysis that smoking and cancer were highly correlated.
- So how do we know that Fisher was wrong? (He was)

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:
  1. **Strength** Causal events should have strong correlation.
  2. **Consistency** Different studies should show similar effect.
  3. **Specificity** A single cause should lead to a single effect.
  4. **Temporality** The effect should occur before the cause.
  5. **Gradient** Greater exposure to cause should correspond to greater size of effect
  6. **Plausibility** A plausible mechanism should exist linking cause and effect.
  7. **Coherence** A cause and effect relationship should not conflict with other known relationships
  8. **Experimental Evidence** A cause and effect relationship should be evident in randomized experiment.
  9. **Analogy** A cause and effect relationship should also be observed in other similar phenomena

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:
    1. **Strength** Causal events should have strong correlation.
    2. **Consistency** Different studies should show similar effect.
    3. **Specificity** A single cause should lead to a single effect.
    4. **Temporality** The effect should occur before the cause.
    5. **Gradient** Greater exposure to cause should correspond to greater size of effect
    6. **Plausibility** A plausible mechanism should exist linking cause and effect.
    7. **Coherence** A cause and effect relationship should not conflict with other known relationships
    8. **Experimental Evidence** A cause and effect relationship should be evident in randomized experiment.
    9. **Analogy** A cause and effect relationship should also be observed in other similar phenomena
- Are these *absolutely* necessary to prove causality?

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:
  1. **Strength** Causal events should have strong correlation.
  2. **Consistency** Different studies should show similar effect.
  3. **Specificity** A single cause should lead to a single effect.
  4. **Temporality** The effect should occur before the cause.
  5. **Gradient** Greater exposure to cause should correspond to greater size of effect
  6. **Plausibility** A plausible mechanism should exist linking cause and effect.
  7. **Coherence** A cause and effect relationship should not conflict with other known relationships
  8. **Experimental Evidence** A cause and effect relationship should be evident in randomized experiment.
  9. **Analogy** A cause and effect relationship should also be observed in other similar phenomena
- Are these *absolutely* necessary to prove causality?
  - No. But they are good guidelines.

Section 3

Introduction to Linear Regression

Overview

"All models are wrong, but some are useful."

— George Box, statistician

Overview

"All models are wrong, but some are useful."

— George Box, statistician

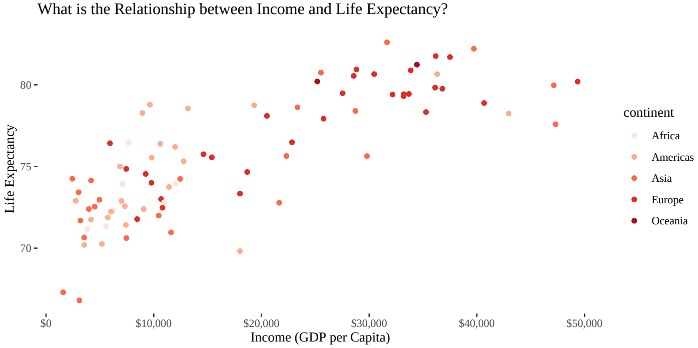
- Linear regression is both an accessible and potent tool in statistical analysis.

Overview

"All models are wrong, but some are useful."

— George Box, statistician

- Linear regression is both an accessible and potent tool in statistical analysis.

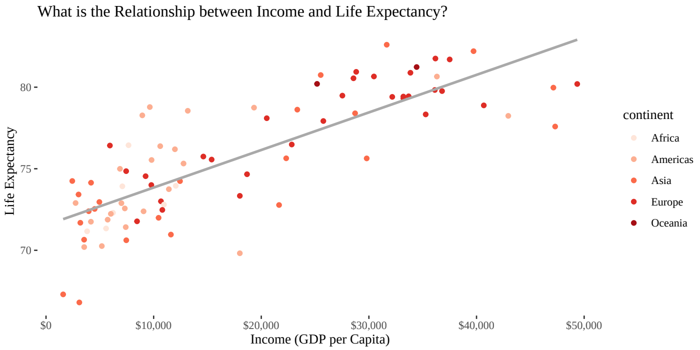


Overview

"All models are wrong, but some are useful."

— George Box, statistician

- Linear regression is both an accessible and potent tool in statistical analysis.



Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

$$Y = \beta_0 + \beta_1 X \qquad \text{with } \beta_0, \beta_1 \text{ fixed constants}$$

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

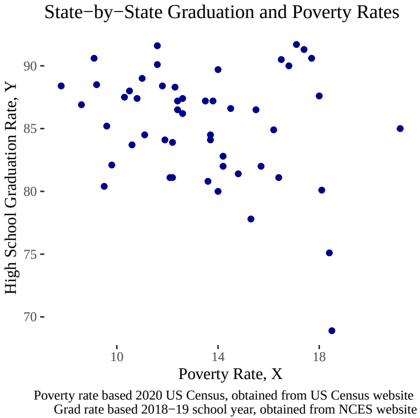
- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

$$Y = \beta_0 + \beta_1 X \qquad \text{with } \beta_0, \beta_1 \text{ fixed constants}$$

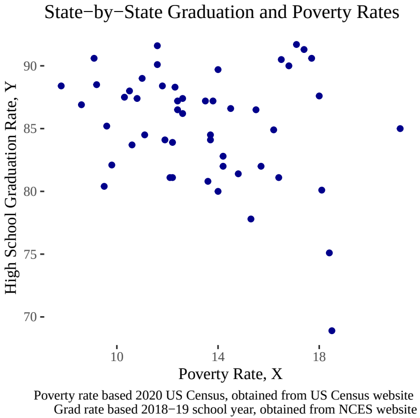
- Of course, in the wild, the observed values of Y will **not** be perfectly predicted by the values of X.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \text{where } \epsilon \text{ is the error}$$

Scatterplots and Linear Relationships I

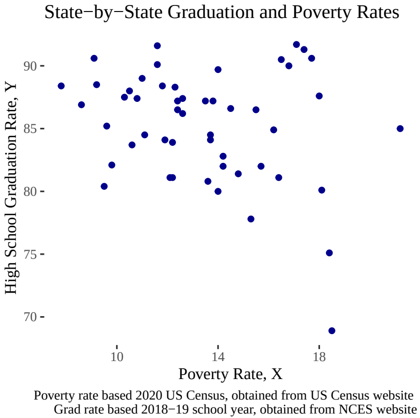


Scatterplots and Linear Relationships I



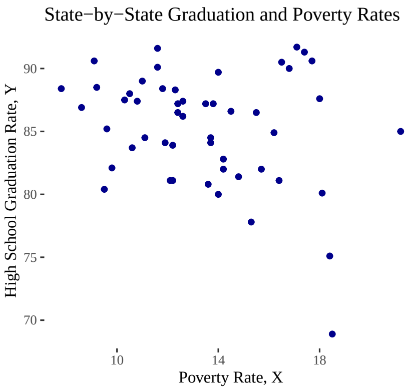
- Explanatory Variable:

Scatterplots and Linear Relationships I



- Explanatory Variable:
  - Poverty Rate (X)

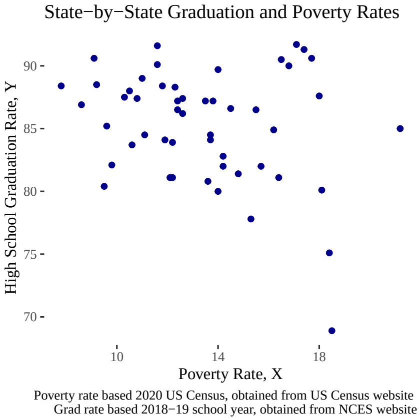
Scatterplots and Linear Relationships I



Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

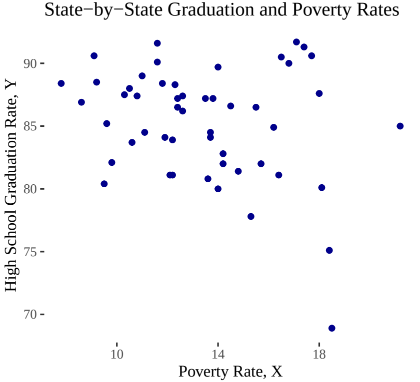
- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:

Scatterplots and Linear Relationships I



- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)

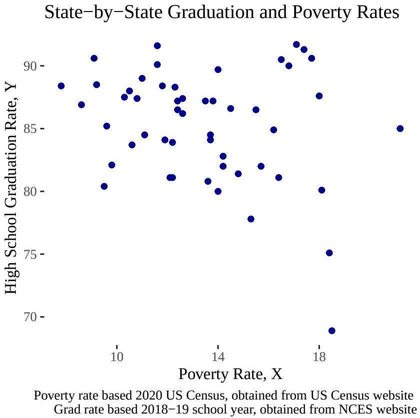
Scatterplots and Linear Relationships I



Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

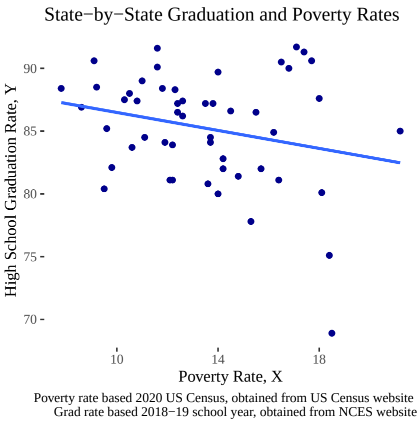
- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)
- Relationship:

Scatterplots and Linear Relationships I



- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)
- Relationship:
  - Linear, negative, moderately strong

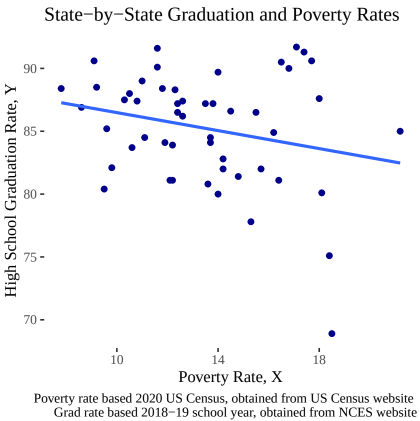
Scatterplots and Linear Relationships II



- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

Scatterplots and Linear Relationships II

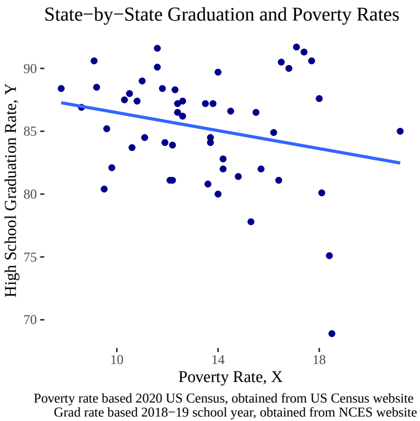


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

- Hat ($\hat{Y}$) indicates this is an estimate of Y

Scatterplots and Linear Relationships II

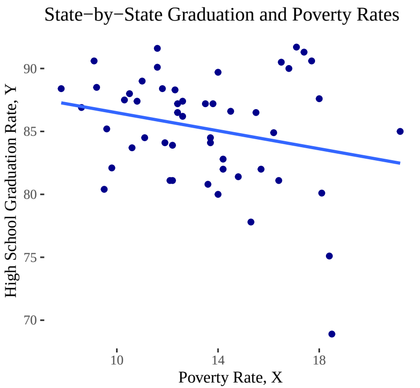


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

- Hat ($\hat{Y}$) indicates this is an estimate of Y
- **Slope** of $\beta_1 = -0.4$ means every 1 unit increase in Poverty corresponds to a 0.4 unit decrease on average in Graduation.

Scatterplots and Linear Relationships II



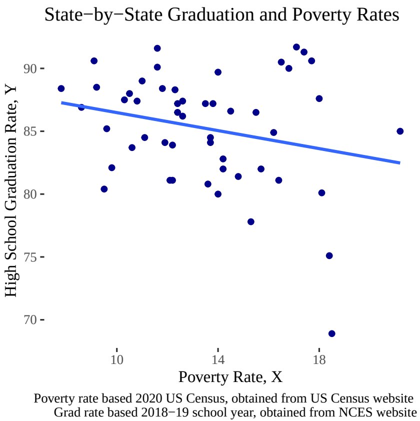
Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

- Hat ($\hat{Y}$) indicates this is an estimate of Y
- **Slope** of $\beta_1 = -0.4$ means every 1 unit increase in Poverty corresponds to a 0.4 unit decrease on average in Graduation.
- **Intercept** of $\beta_0 = 90$ means model predicts graduation rate of 90% when poverty rate is 0%.

Scatterplots and Linear Relationships III

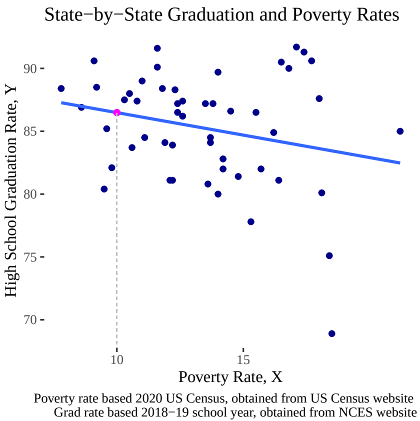


- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 10%?

Scatterplots and Linear Relationships III

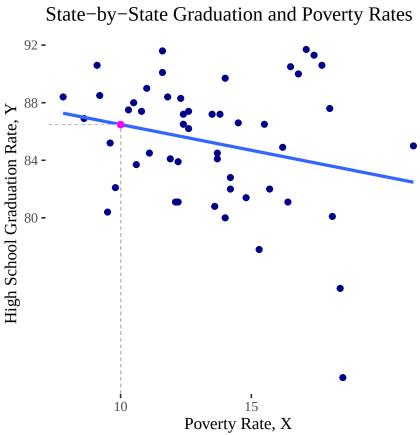


- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 10%?

Scatterplots and Linear Relationships III



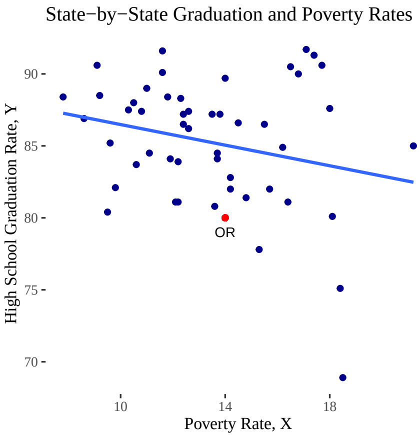
- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 7%?

$$\hat{Y} = 90 - 0.4 \cdot 10 = 86$$

Scatterplots and Linear Relationships IV



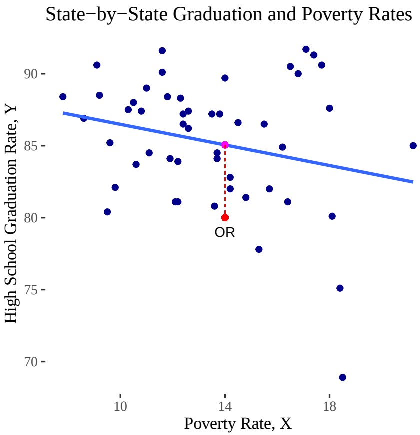
- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- Oregon has a poverty rate of 14. What does the model predict is Oregon's graduation rate?

$$\hat{Y} = 90 - 0.4 \cdot 14 = 84.4$$

Scatterplots and Linear Relationships IV



- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- Oregon has a poverty rate of 14. What does the model predict is Oregon's graduation rate?

$$\hat{Y} = 90 - 0.4 \cdot 14 = 84.4$$

But Oregon's actual graduation rate is 80

Residuals

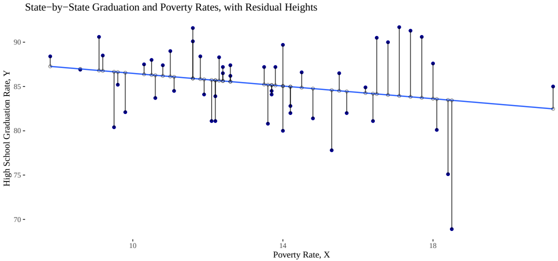
- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

$$e_i = Y_i - \hat{Y}_i$$

Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

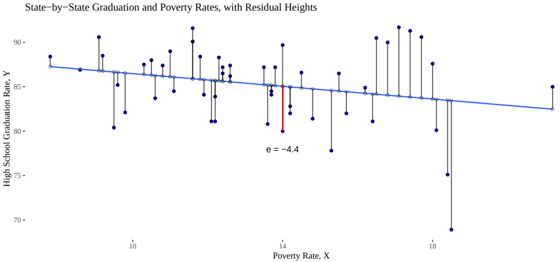
$$e_i = Y_i - \hat{Y}_i$$



Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

$$e_i = Y_i - \hat{Y}_i$$

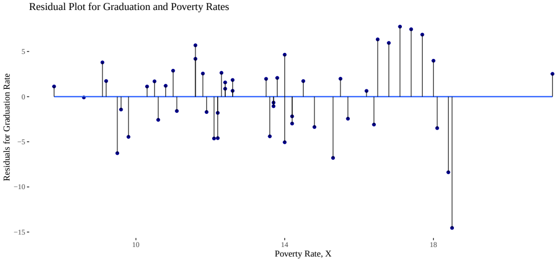


-Oregon's residual is

$$e = Y - \hat{Y} = 80 - 84.4 = -4.4$$

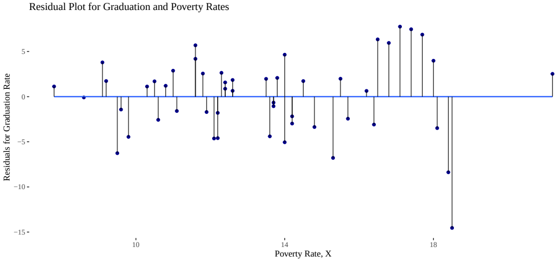
Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



Residual Plot

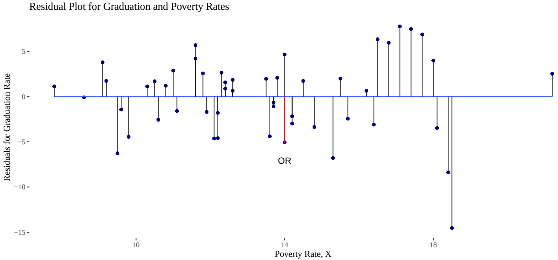
- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.

Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.